

# The hourglass and the early conservation models — co-existing evolutionary patterns in vertebrate development

Barbara Piasecka<sup>1,2,4</sup>, Pawel Lichocki<sup>3</sup>, Sven Bergmann<sup>2,4,#</sup> Marc Robinson-Rechavi<sup>1,4,#,\*</sup>

**1** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

**2** Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

**3** Laboratory of Intelligent Systems, EPFL, Lausanne, Switzerland

**4** Swiss Institute of Bioinformatics, Lausanne, Switzerland

**#** These authors contributed equally to this work.

**\*** Corresponding author

**Address:** Marc Robinson-Rechavi, Biophore, UNIL, 1015 Lausanne, Switzerland

**E-mail:** marc.robinson-rechavi@unil.ch

**Tel:** +41 21 692 42 20

**Running title:** Evolutionary patterns in vertebrate development

**Keywords:** modules, vertebrate genome evolution, developmental constraints, hourglass model

## Abstract

Developmental constraints have been postulated to limit the space of feasible phenotypes and thus shape animal evolution. These constraints have been suggested to be the strongest during either early or mid-embryogenesis, which corresponds to the early conservation model or the hourglass model, respectively. Conflicting results have been reported, but in recent studies of vertebrate transcriptomes the hourglass model has been favored. Studies usually report descriptive statistics calculated for all genes over all developmental time points. This introduces dependencies between the sets of compared genes, and may lead to biased results. Here we overcome this problem using an alternative modular analysis. We used the Iterative Signature Algorithm to identify distinct modules of genes co-expressed specifically in consecutive stages of zebrafish development. We then performed a detailed comparison of several gene properties between modules, allowing for a less biased and more powerful analysis. Notably, our analysis corroborated the hourglass pattern only at the regulatory level, with sequences of regulatory regions being most conserved for genes expressed in mid-development, but not at the level of gene sequence, age or expression, in contrast to some previous studies. The early conservation model was supported with gene duplication and birth that were the most rare for genes expressed in early development. Finally, for all gene properties we observed the least conservation for genes expressed in late development or adult, consistent with both models. Overall, with the modular approach, we showed that different levels of molecular evolution follow different patterns of developmental constraints, and thus neither model is exclusively valid.

## Introduction

Developmental constraints have been suggested to play an important role in shaping the evolution of embryonic development in animals. Briefly, the concept of developmental constraints assumes that the scope of developmental mechanisms limits the set of phenotypes that may evolve. Thus, morphological similarities between embryos of different species could reflect these underlying constraints (Poe and Wake 2004). Two main models of embryonic developmental constraints have been put forward. The *early conservation* model predicts that the highest developmental constraints occur at the beginning of embryogenesis. This corresponds to von Baer’s third law (von Baer 1828), postulating that embryos of different species progressively diverge from one another during ontogeny. However, in modern times, the

highest morphological similarity between embryos of different species was observed in the *phylotypic stage* (i.e., mid-embryogenesis) (Seidel 1960; Sander 1983; Elinson 1987). Consequently, Duboule (1994) and Raff (1996) proposed the so-called *hourglass* model, which has since become widely accepted (see, e.g., Prud'homme and Gompel 2010; Kalinka and Tomancak 2012). It predicts the highest developmental constraints during mid-embryogenesis.

At the genomic level, the hourglass model was originally linked to the expression of Hox genes in vertebrates (Duboule 1994). More recently, the emphasis has shifted to the relation, if any, between developmental constraints and the evolution and function of the genome (reviewed in Kalinka and Tomancak 2012). Different studies have reported several characteristics supporting the hourglass model in vertebrates on the genomic level, e.g.: higher protein sequence similarity (Hazkani-Covo et al. 2005), higher expression conservation (Irie and Kuratani 2011), and older age (Domazet-Lošo and Tautz 2010) of genes expressed in the mid-development when compared to the genes expressed early or late in the development. Very recently, the hourglass model was proposed also for plants embryogenesis, for gene age and for sequence conservation (Quint et al. 2012). However, some of these results do not hold out under a detailed analyses (see Box 1 and Supplementary Materials). For example, applying a standard log-transformation (McDonald 2009; Speed 2000) to microarray signal intensities used in Domazet-Lošo and Tautz (2010) changes the reported pattern such that it no longer supports the hourglass model (figure 1). Moreover, other studies have also found genetic patterns supporting an early conservation model (Roux and Robinson-Rechavi 2008; Comte et al. 2010).

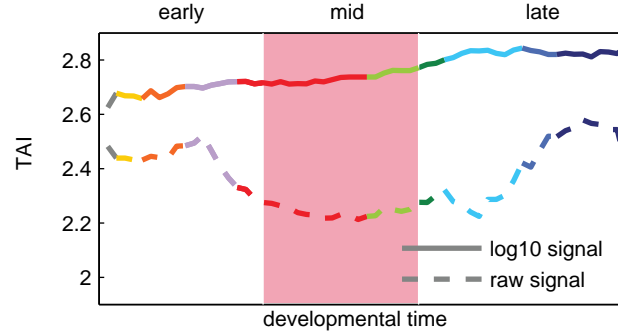
In most of the studies of developmental constraints the authors compared descriptive statistics of all genes across all developmental time-points (e.g., median expression (Roux and Robinson-Rechavi 2008), weighted mean age (Domazet-Lošo and Tautz 2010), mean expression correlation (Irie and Kuratani 2011)). Such an approach introduces dependencies between the sets of genes which are compared, and consequently can produce results biased by genes expressed at many time-points. For example, house-keeping genes contribute to the average gene expression at all time points, and hence dilute trends. To overcome this essential problem, we have used a *modularization* approach, which we applied to the recently published transcriptome data of zebrafish development (Domazet-Lošo and Tautz 2010). We decomposed the genes into independent sets, i.e., *modules*, that contained genes overexpressed solely in one of seven developmental stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile and adult. This decomposition allowed us to compare only sets of genes that have specific functions

during embryonic development. For each of the seven modules, we studied five properties of its genes: 1) gene sequence conservation, 2) gene age, 3) gene expression conservation, 4) gene orthology relationships, and 5) regulatory elements conservation.

Here, we show that different levels of molecular evolution follow different patterns of developmental constraints. First, the regulatory elements are most conserved for transcription factors expressed at mid-development, consistent with the hourglass model. Contrary to what has been reported previously (Hazkani-Covo et al. 2005; Domazet-Lošo and Tautz 2010; Irie and Kuratani 2011), we did not detect the hourglass pattern for gene sequence, age and expression. Second, constraints on gene duplication and on new gene introduction are the strongest in early development, supporting the early conservation model (consistent with Roux and Robinson-Rechavi 2008). Finally, all gene properties displayed the least conservation in late development and adult, which is in agreement with both models of developmental constraints.

### Box 1: Transcriptome Age Index

Recent results of Domazet-Lošo and Tautz (2010) suggest that the oldest transcriptome set is expressed at the phylotypic stage, and that younger sets are expressed during early and late development, which supports the hourglass model. To study the relationship between gene expression, ontogeny and phylogeny, the authors proposed a measure called the “transcriptome age index”, or TAI. The TAI was defined as the mean of the phylogenetic ranks (“phylostrata”) across genes, weighted by their microarray signal intensity values at each developmental stage. Note that the microarray signal intensity values used in Domazet-Lošo and Tautz (2010) displayed a log-normal distribution and spanned from 1 to  $10^5$  (Supplementary figure S1). Using these values to calculate TAI made the weights of phylogenetic ranks differ by five orders of magnitude between lowly and highly expressed genes. Consequently, only the most expressed genes (Supplementary figure S2), and potentially outliers (Supplementary figure S3), contributed to the hourglass pattern discovered with TAI. We found that applying a standard log-transformation to the intensity values changes the pattern, which then indicates older genes being expressed preferentially in early development (figure 1). The use of log-transformed data for microarray intensities is generally encouraged (McDonald 2009; Speed 2000) because it keeps the biological signal, while removing dependency between variance and intensity of the analyzed signals. We present a more detailed re-analysis of the study of Domazet-Lošo and Tautz (2010) in Supplementary Materials.



**Figure 1: Transcriptome age index (TAI) using raw and log-transformed expression signal intensities.** A higher TAI value implies that evolutionary younger genes are preferentially expressed at the corresponding time-point. The pink shaded area indicates the phylotypic stage. Colors of the curves reflect the main developmental periods and correspond to the colors used in Domazet-Lošo and Tautz (2010).

## Results

### Modules

Our goal was to analyze the developmental constraints acting on different gene properties. To this end we identified and analyzed groups of genes co-expressed during distinct developmental stages. We applied the Iterative Signature Algorithm (ISA) (Bergmann et al. 2003; Ihmels et al. 2004) to the zebrafish expression data published by Domazet-Lošo and Tautz (2010), which measured the dynamics of the transcriptome during development with a resolution of 60 time points. The ISA is a modularization algorithm that finds genes with similar expression profiles and groups them into so-called transcription modules. In order to detect modules of genes with specific expression during the zebrafish development, we initialized the ISA with seven idealized expression profiles that corresponded to successive developmental stages (see Supplementary Materials and Supplementary figure S8).

We obtained seven modules, each containing genes overexpressed during one of the following developmental stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile and adult (figure 2). Overall, the modules covered the entire development. The phylotypic stage in which the hourglass model predicts the highest evolutionary constraints corresponds to the segmentation and pharyngula modules. We will refer to these two modules as phylotypic modules. The cleavage/blastula and gastrula modules will be referred to as early modules, and larva, juvenile and adult modules as late modules.

The adjacent modules partially overlapped in their gene content. In order to allow for unbiased cross-module comparisons, genes belonging to two modules were kept in the one with the highest ISA gene score (see Methods); this concerned 534 genes in total. The seven modules, i.e., cleavage/blastula, gastrula, pharyngula, segmentation, larva, juvenile and adult, contained 444, 820, 487, 414, 415, 290 and 207 genes, respectively (see Methods for the lists of the genes). Overall, 3077 different genes were present in these modules, which implies a significant reduction of the number of genes being analyzed in comparison to the original data (14 293 genes on the microarray). In particular, the ISA removed the bias related to the genes expressed uniformly across development (i.e., housekeeping genes).

### Functional annotation

We verified the function of genes in modules detected by the ISA by comparing them to relevant known lists of genes. We found that the cleavage/blastula module was significantly enriched in maternal genes

identified in (Aanes et al. 2011) (hypergeometric test,  $p = 0.01$ , see Methods for details of this, and all other statistical tests), and the gastrula module was highly significantly enriched in post-midblastula transition (post-MBT) genes identified in Aanes et al. (2011) (hypergeometric test,  $p = 2.8 \times 10^{-18}$ ). We confirmed the relevance of the pharyngula and segmentation modules by verifying that they were enriched in Hox genes, which is consistent with their role in mid-development (Krumlauf 1994) (hypergeometric test,  $p = 5.6 \times 10^{-16}$  and  $2.9 \times 10^{-4}$ , respectively). We did not have any gold standard for genes expressed at the late stages of development. However, since the early and phylotypic modules were enriched in genes with relevant functions, we are confident that the same is true for the late modules.

Moreover, GO enrichment analysis confirmed that genes from the modules were enriched in functions relevant to the respective developmental stages. For example, the cleavage/blastula module was enriched in genes involved in protein phosphorylation and dephosphorylation processes, which is consistent with kinase-dependent control of cell cycle and regulation of mid-blastula transition (MBT) in vertebrates (Hartley et al. 1996; Yarden and Geiger 1996). The pharyngula module was enriched in genes associated with cell differentiation, and anatomical structure development. Finally, the adult module was enriched in genes involved in responses to environment, although not significantly (Supplementary table S2).

## Sequence conservation

We checked whether the sequences of genes from different modules evolved under different selective pressure. To this end, we calculated the non-synonymous to synonymous substitution ratios ( $d_N/d_S$ ) for genes in the modules and asked if the ratio was significantly lower for any of them. With the early conservation model, we would expect the lowest  $d_N/d_S$  values for genes from early modules. Whereas with the hourglass model, we would expect the lowest  $d_N/d_S$  values for genes from the phylotypic modules. In the first five modules, covering whole embryonic development from zygote to larva, the median  $d_N/d_S$  was lower than the median  $d_N/d_S$  for all genes, but the difference was significant only for the larva module (figure 3A, randomization test,  $p < 7 \times 10^{-4}$ ). In the juvenile module, the median  $d_N/d_S$  was higher than the median  $d_N/d_S$  for all genes, but the difference was not significant. In the adult module, the median  $d_N/d_S$  was significantly higher than the median  $d_N/d_S$  for all genes (randomization test,  $p = 4.2 \times 10^{-3}$ ).

These results were consistent with the study by Roux and Robinson-Rechavi (2008), who also reported equally low  $d_N/d_S$  values during the entire zebrafish embryogenesis, and a small increase in mid-larva,

juvenile and adult. In contrast, Hazkani-Covo et al. (2005) reported an hourglass pattern for protein distance between mouse and human genes expressed during development. However, the trend was not significant. In Roux and Robinson-Rechavi (2008) some evidence for early conservation was reported in mouse. Projecting the genes from zebrafish modules to mouse-human orthologs, we found equal conservation across development (Supplementary figure S9). Overall, data analyses support similar evolutionary constraints on sequences of genes expressed during whole embryogenesis of zebrafish, while for mouse more developmental data is needed to be conclusive.

## Gene age

The differences in age of genes expressed during different stages of the development have been suggested to be a good indicator of evolutionary constraints (Irie and Sehara-Fujisawa 2007; Domazet-Lošo and Tautz 2010). Thus, we investigated the age of genes belonging to different modules. We dated each gene by its first appearance in the phylogeny and assigned it to one of the five age groups: 1) Fungi/Metazoa, 2) Bilateria, 3) Coelomata+Chordata, 4) Euteleostomi and 5) Clupeocephala+*Danio rerio*. Next, for each module we calculated the age distribution of its genes, i.e., the number of genes belonging to each age group, and compared it with the age distribution of all genes.

For all but the cleavage/blastula module we detected significant age variations (chi-square goodness of fit test, all  $p < 1.3 \times 10^{-5}$ ), which differed across modules. The oldest genes were overrepresented in the gastrula module, the Bilateria genes were overrepresented in the phylotypic modules, and the youngest genes were overrepresented in the late modules (figure 3B). In contrast, Domazet-Lošo and Tautz (2010) reported that genes expressed in early and late development tend to be younger than genes expressed in mid-development, supporting the hourglass model. Yet, that result does not hold for log-transformed gene expression levels (Box 1), and is not recovered with measures of gene age other than the transcriptome age index (see Supplementary Materials and Supplementary figure S6). With the modular approach we observed that the age of expressed genes decreased throughout ontogeny. This pattern suggests that the oldest evolutionary stages tend to express the oldest genes.

## Gene family size

Both gene duplication and gene loss can impact phenotypic evolution (Ohno et al. 1970; Zhang 2003; Nei 2007; Wang et al. 2006; Demuth and Hahn 2009). The outcome of these events can be summarized



by the resulting gene family size. Consequently, constrained developmental stages should display less changes in gene family size than other stages. To test this hypothesis, for each zebrafish module we calculated the number of its genes that were in 1) one-to-one, 2) one-to-many, 3) many-to-many, and 4) no orthology relation to mouse genes (i.e., no ortholog detectable by the criteria used in Ensembl Compara; Vilella et al. 2009).

We compared the observed distributions with the distribution of the ortholog relationships for all genes. We detected significant variations of the ortholog relationship for the cleavage/blastula module and for all three late modules (chi-square goodness of fit test, all  $p < 9 \times 10^{-5}$ ). Moreover, the pattern of variation itself differed across different modules. The number of one-to-one orthologs decreased throughout development, and was significantly higher than expected only in the cleavage/blastula module (figure 3C). In contrast, the number of genes with no orthologous relationship increased throughout development. It was significantly higher than expected only in the juvenile and adult modules (figure 3C), consistent with the excess of “young” genes. A similar pattern was observed for many-to-many orthologs. Finally, the number of one-to-many orthologs was higher than expected only in the larva module, and did not differ from expectation in all other modules.

These results were consistent with Roux and Robinson-Rechavi (2008) in which the genes retained in duplicates after the fish-specific whole genome duplication were reported to have low expression early in the development. Here, we recovered an analogous pattern with the modular approach, showing that the genes expressed early in the development are retained in duplicates less often than genes expressed later. Note that our observation is not limited to whole genome duplication. In addition, we detected the highest number of novel genes amongst genes expressed late in the development.

## Expression conservation

Changes in gene expression are one of the main sources of morphological variation (King and Wilson 1975; Preuss et al. 2004; Carroll 2005). The developmental constraints on gene expression might differ from those on the gene sequence (Jordan et al. 2004; Yanai et al. 2004; Jordan et al. 2005). Thus, for each module, we compared the mean expression profile of its genes with the mean expression profile of their one-to-one orthologs in mouse. We used two different data sets (Wang et al. 2004; Irie and Kuratani 2011) with expression values of mouse genes during the development. The use of two data sets was necessary, because there does not exist a single experiment covering the entire mouse development. The

incompatibility of the two microarrays impaired the statistical strength of the analysis. For this reasons the results reported here should be regarded rather as qualitative than quantitative.

Since homology cannot be defined for individual developmental stages between zebrafish and mouse, we first mapped every time point to its broad metastage defined in Bgee (Bastian et al. 2008) (figure 4). Next, we calculated the mean expression level in every metastage. This resulted in six expression values for each gene during the development of mouse and zebrafish: zygote, cleavage, blastula, neurula, organogenesis, and post-embryonic stage. Note that the mouse microarrays did not cover the gastrula stage at all. For each module we calculated the Pearson’s correlation between the mean expression of its genes and their mouse orthologs across the six metastages. For the cleavage/blastula module no correlation was detected, probably due to the incompatibility of the two mouse microarrays. For other modules the correlation was positive (figure 3D), however due to the low number of data points in the analysis, no correlation values were significant (all  $p > 0.01$ ).

These results stood in contrast with the report by Irie and Kuratani (2011) who showed the highest conservation of gene expression in mid-development. However, a re-analysis of their data suggested that this observation was not significant (see Supplementary Materials and Supplementary figure S7). Also, both their and our studies shared problems related to the use of two data sets from different sources to cover mouse development. This and the lack of a straightforward homology between ontogenies of different species make it difficult to conclude on the conservation of gene expression during vertebrate development.

## Regulatory regions

The *cis*-regulatory hypothesis asserts that most morphological evolution is due to changes in *cis*-regulatory sequences (Stern 2000; Wray 2007; Carroll 2008). A reasonable prediction of this hypothesis is slower *cis*-element turnover in morphologically conserved developmental periods. We examined the presence of highly conserved non-coding elements (HCNEs; Engström et al. 2008) and of transposon-free regions (TFRs; Simons et al. 2007) in the proximity of genes from each module. In the analysis of HCNEs, we counted their number between zebrafish and mouse (detected with 70% identity) in regions of 500 base pairs upstream from the transcription start site. We found that only genes from the phylotypic modules were significantly enriched in HCNEs (hypergeometric test,  $p = 8 \times 10^{-6}$ , and  $p = 1.1 \times 10^{-4}$  for segmentation and pharyngula modules, respectively). We tested the sensitivity of the results by changing the

analyzed regions' length to 200 and 1000 base pairs upstream from the transcription start site, by looking for HCNEs in introns, and using HCNEs detected with identity of 90%. In all cases, we obtained similar results (see Supplementary table S1). In the analysis of TFRs, we counted the number of genes from each module that have been associated with TFRs in zebrafish. Importantly, these TFRs were reported to be conserved between vertebrates as distant as zebrafish and human. We found that only genes from the pharyngula module were significantly enriched in TFRs (hypergeometric test,  $p = 5.7 \times 10^{-7}$ ).

The highly conserved non-coding elements and transposon-free regions are often associated with developmental regulatory genes, and with transcription factors (TFs) in particular (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Engström et al. 2008; Simons et al. 2007). In order to confirm this association, we calculated the fractions of genes with HCNEs or with TFRs in their proximity. We observed that for both features this fraction was higher for TFs than for all genes. Importantly, we observed that only the phylotypic modules were enriched in TFs (figure 3E). This partially explained the enrichment in HCNEs and TFRs for genes expressed in mid-development. In addition, HCNEs were more often present in the proximity of TFs from the pharyngula module than in the proximity of TFs in general (figure 3E; 8.8% of TFs from the pharyngula module had at least one HCNE in their proximity, and only 3.7% of all TFs had at least one HCNEs in their proximity). Also TFRs were more often present in the proximity of TFs from the phylotypic modules than in the proximity of TFs in general (figure 3E; 31% and 45% of TFs from the segmentation and pharyngula modules, respectively, had TFRs in their proximity, and only 26% of all TFs had TFRs in their proximity). Consequently, the enrichment in HCNEs and TFRs for genes expressed in the phylotypic stage seems to be related to the regulation of developmental processes. Interestingly, only few Hox genes from phylotypic modules were associated with HCNEs (four Hox genes from segmentation module), and with TFRs (six Hox genes from segmentation module, and one Hox gene from pharyngula module).

In addition, we checked for genes that preserved their specific ancestral order in the genome across metazoans (so called conserved ancestral microsyntenic pairs; Irimia et al. 2012) and are known to be involved in the regulation of development. We found that they were slightly overrepresented in the segmentation module, but only at the limit of statistical significance (see Supplementary Materials).

Finally, we checked for core developmental genes in each module (see Vavouri et al. 2007 for the list of genes). These genes are known to be involved in the regulation of development, and to have highly conserved regulatory regions within different taxa, including, nematodes, insects and vertebrates

(Vavouri et al. 2007). We detected a significant enrichment in these genes only in the pharyngula module (20 core genes; hypergeometric test,  $p = 6.9 \times 10^{-19}$ ), supporting the hourglass model.

## Discussion

Our goal was to study developmental constraints acting on various gene properties. To this end we identified distinct sets of genes with time-specific expression in zebrafish development, i.e., genes expressed in one of the seven consecutive stages: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile and adult. Overall, we analyzed and compared five gene characteristics, namely the conservation of gene sequence, gene expression, and regulatory elements, as well as age and orthology relationships.

Several features do not show any significant pattern over embryonic development, often in contradiction to previous reports. There is notably no evidence for change in selective pressure acting on sequences of protein-coding genes (i.e.,  $d_N/d_S$ ) over development (in contrast to Hazkani-Covo et al. 2005). Unfortunately, the available data does not allow a strong conclusion concerning the conservation of expression (in contrast to Irie and Kuratani 2011), despite the probable importance of this feature in the evolution of development. In this respect, the situation in vertebrates stands in contrast to the relatively clear results in flies (Kalinka et al. 2010), where the evolution of expression has been shown to be most constrained in mid-development.

Gene orthology relations support the early conservation model. We show that early stages are less prone to tolerate both gene duplication (consistent with Roux and Robinson-Rechavi 2008) and gene introduction. The interpretation of transcriptome age is less straightforward. Our observations suggest that the oldest evolutionary stages tend to express of the oldest genes. It is possible that early stages are evolutionarily oldest, and that this is why they are enriched in oldest genes. Consequently, it is the presence of young genes in a module that would mark relaxed developmental constraints during the corresponding stage. However, neither early nor phylotypic modules are enriched in young genes (Euteleostomi and Clupeocephala+*Danio rerio*), which suggests similar developmental constraints in early and mid-ontogeny. In any case, we do not find any support for the hypothesis that the phylotypic stage would be characterized by the oldest transcriptome (in contrast to Domazet-Lošo and Tautz 2010).

While the modularization approach does not support several previous hypotheses of genomic traces of the phylotypic stage, it allows us to distinguish a strong signal of conservation of gene regulation in

mid-development. While this had not yet been reported in genomic studies, it is consistent with early descriptions of the phylotypic stage as characterized by Hox genes body patterning activity (Duboule 1994). Of note, the patterns that we observe are robust to the removal of Hox genes, so they are more general than this original observation. We observed an excess of HCNEs only for genes expressed in the pharyngula module, and an excess of TFRs only for genes expressed in the phylotypic modules. The enrichment in HCNEs and TFRs has been related to developmental regulatory genes, and to transcription factors (TFs) in particular (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Engström et al. 2008). Indeed, we observed that more TFs were expressed in mid-development than in other stages. Also, we showed that a significant proportion of TFs expressed in mid-development had conserved regulatory regions (i.e., HCNEs and TFRs), in contrast to TFs expressed early or late. Consequently, the enrichment in HCNEs and TFRs for genes expressed in mid-development can be explained by both a higher number of TFs and a higher number of HCNEs and TFRs for these TFs, than for genes expressed earlier or later. Moreover, the pharyngula module was associated with core developmental genes. Overall, these results suggest that mid-developmental processes have extremely high conservation of regulation. This conservation could translate into observed common traits of the phylum expressed at the phenotypic level during mid-development. In addition, core developmental genes are known to be present in different taxa (e.g., nematodes, insects and vertebrates), in each of which they have a conserved regulation that evolved in parallel (Vavouri et al. 2007). This could explain why the phylotypic stage is observed not only in vertebrates (Kimmel et al. 1995), but also in other phyla, e.g., in arthropods (Sander 1983; Kalinka et al. 2010).

Finally, for all of the features which we have considered there is at least some trend towards weaker evolutionary constraints in the latest stages:  $d_N/d_S$  is significantly higher in adults; correlation of expression is lowest for maternal, larval and adult genes; young genes and genes with duplications in fishes or other vertebrates are overrepresented in late modules; and genes expressed in juveniles and adults have the less HCNEs and TFRs. Although not all of these trends are significant, no feature shows stronger conservation in late development or adult. Thus, while different aspects of gene evolution show constraints at different times of development, there appears to be a generally faster evolution of all aspects of larval, juvenile and adult genes. Whether this is due to lower constraints (i.e., less purifying selection) or to stronger involvement in adaptation (i.e., more diversifying selection), remains an open question.

In summary, we studied evidence for, or against, any particular pattern of developmental constraints

by considering sets of genes with time-specific expression patterns. Comparing such independent sets of genes with a clear function during embryogenesis resulted in cleaner and more fine-grained characterization of evolutionary patterns than previously reported. Notably, we showed that different levels of molecular evolution follow different patterns of developmental constraints. The sequence of regulatory regions is most conserved for genes expressed in mid-development, consistent with the hourglass model. Gene duplication and new gene introduction is most constrained during early development, supporting the early conservation model. Whereas, all gene properties coherently show the least conservation for the latest stages, consistent with both the early conservation and the hourglass models.

## Methods

### Gene expression data

Microarray data of zebrafish development were downloaded from NCBI's Gene Expression Omnibus (Edgar et al. 2002) (GSE24616). This study was performed on the Agilent Zebrafish (V2) Gene Expression Microarray. In total, expression profiles for 60 developmental stages (from unfertilized egg to adults stages) were measured. The last ten stages (55 days - 1 year 6 months) were measured separately for male and female. Two replicates were made per time point, resulting in  $(50 + 2 \times 10) \times 2 = 140$  microarrays in total. For each microarray, values of `gProcessedSignal` were  $\log_{10}$  transformed and normalized as follows. Separately for each replicate, we equalized the expression signals between microarrays using the spike-ins reference, to account for different amounts of RNA present throughout development. To this aim, we first quantile normalized the expression signal of all spike-ins from all microarrays. Then, for each spike-in level we took the median value of expression signal before and after quantile normalization. This resulted in 10 pairs of expression signals (original signal vs. normalized signal). With linear interpolation between these points, we obtained a piecewise linear curve that defined a mapping from original to normalized expression signals, which we used to equalize the expression signals from all microarrays. This was done by projecting each expression signal onto the piecewise linear curve and calculating the corresponding normalized value. Finally, we quantile normalized the data within replicates and computed the mean value for each gene within replicates. Expression values measured separately for males and females were averaged for each time point.

Microarray data of mouse development were downloaded from Array Express (E-MEXP-51 and E-

MTAB-368). The E-MEXP-51 study was performed on (C57BL/6×CBA)F1 mice using Affymetrix GeneChip Murine Genome U74Av2. In total, expression profiles for 10 early developmental stages (zygote, early 2-cell, mid 2-cell, late 2-cell, 4 cell, 8 cell, 16 cell, early blastocyst, mid-blastocyst, late blastocyst) were measured. 2-4 replicates were made per time point. The data were normalized using gcRMA package.

The E-MTAB-368 study was performed on C57BL/6 mice using Affymetrix GeneChip Mouse Genome 430 2.0. In total, expression profiles for 8 mid and late developmental stages (E7.5, E8.5, E9.5, E10.5, E12.5, E14.5, E16.5, E18.5) were measured. 2-3 replicates were made per time point. The data were normalized using gcRMA package.

## Mapping probe sets to Ensembl genes

Agilent probe sets were mapped to their corresponding zebrafish genes (Ensembl release 63; Hubbard et al. 2009) using BioMart (Smedley et al. 2009). Probe sets which did not map unambiguously to an Ensembl gene were excluded from the analysis. A total of 19 049 probe sets corresponding to 14 293 zebrafish genes were taken into account in our analysis.

Affymetrix probe sets were mapped to their corresponding mouse genes (Ensembl release 63; Hubbard et al. 2009) using BioMart (Smedley et al. 2009). Probe sets which did not map unambiguously to an Ensembl gene were excluded from the analysis. For genes that were mapped by several probe sets we used the signal averaged across the probe sets. A total of 2883 mouse genes mapped by probe sets present on both mouse microarrays were taken into account in the gene expression analysis.

## Iterative Signature Algorithm (ISA)

The ISA identifies modules by an iterative procedure. A detailed description of the algorithm in the general case is given in (Bergmann et al. 2003) (see also [http://www2.unil.ch/cbg/homepage/downloads/ISA\\_tutorial.pdf](http://www2.unil.ch/cbg/homepage/downloads/ISA_tutorial.pdf)). In this specific study, the algorithm was initialized with seven candidate seeds, each consisting of one artificial expression profile corresponding to one of the zebrafish developmental stages (see Supplementary Materials for details). Next, these seeds were refined through iterations by adding or removing genes and developmental time points until the processes converge to stable sets, which are referred to as (transcription) modules. Each developmental time point and gene received a score indicating their membership (if non-zero) and contribution to a given module. The closest the score for a gene or developmental time

point was to one, the stronger the association between the gene/developmental time point and the rest of the module.

The ISA was run twice with the following sets of thresholds: 1)  $t_g = 1.8$  and  $t_c = 1.2$ , and 2)  $t_g = 1.8$  and  $t_c = 1.4$ , for genes and developmental time points, respectively. We obtained the pharyngula module only in the case of  $t_c = 1.2$ , and all other modules with both  $t_c = 1.2$  and  $t_c = 1.4$ . All the modules contained their corresponding idealized profile. For further analysis, we kept a single module per developmental stage. From the pair of modules, we chose the one in which the idealized profile had a higher gene score. Overall, segmentation, pharyngula and juvenile modules were obtained with  $t_c = 1.2$ , and cleavage/blastula, gastrula, larva, and adult modules were obtained with  $t_c = 1.4$ .

## Genes belonging to the modules

For the list of zebrafish ensembl genes belonging to the seven modules please refer to the following links.

Cleavage/blastula module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module1.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module1.txt)

Gastrula module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module2.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module2.txt)

Segmentation module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module3.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module3.txt)

Pharyngula module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module4.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module4.txt)

Larva module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module5.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module5.txt)

Juvenile module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module6.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module6.txt)

Adult module: [https://dl.dropbox.com/u/32680149/Genes\\_in\\_modules/Module7.txt](https://dl.dropbox.com/u/32680149/Genes_in_modules/Module7.txt)

## GO enrichment analysis

Gene ontology (GO) association for all genes mapped by zebrafish probe sets were downloaded from Ensembl release 63 (Hubbard et al. 2009), using BioMart (Smedley et al. 2009). GO enrichment was tested by Fisher’s exact test, using the Bioconductor package topGO (Alexa et al. 2006) version 2.2.0. The reference set consisted of all Ensembl genes mapped by probe sets of the microarray used. The “elim” algorithm of topGO was used to eliminate the (tree-like) hierarchical dependency of the GO terms. To correct for multiple testing the Bonferroni correction was applied. For every module GO categories with corrected P-value lower than 0.01 were reported, if less than ten GO categories were significant we reported the top ten (see Supplementary table S2).



## Gene sequence analysis

The orthology relationships, and the values of  $d_N$  (number of non-synonymous substitutions per non-synonymous site) and  $d_S$  (number of synonymous substitutions per synonymous site) were obtained from Ensembl version 63 (Hubbard et al. 2009). We retrieved zebrafish genes with one-to-one orthologs in *Tetraodon nigroviridis* and *Takifugu rubripes* (the estimated divergence time is 32 million years ago (MYA) between the two pufferfish species and 150 MYA with *Danio rerio* (Benton and Donoghue 2007)) and the pairwise  $d_N$  and  $d_S$  between *Tetraodon* and *Takifugu* using Biomart (Smedley et al. 2009). We used the set of 7854 genes having  $d_N$  and  $d_S$  for Tetraodon-Fugu, and having the expression measured on the zebrafish microarray. For every module we calculated the median  $d_N/d_S$  ratio of its  $k$  genes, where  $k$  was the number of genes having one-to-one relationship with *Tetraodon* and *Fugu* genes. Next, we generated 10 000 sets of  $k$  randomly chosen genes. For each set we calculated the median  $d_N/d_S$  ratio. Thus, we constructed a sampling distribution of the median  $d_N/d_S$  values for a set of  $k$  genes. Then we calculated the probability that the median  $d_N/d_S$  of the original module was sampled from the constructed distribution. It allowed us to assess if the observed median  $d_N/d_S$  ratio was significantly different from the expected median value. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level. We repeated the same procedure for mouse-human genes (see Supplementary Materials).

## Gene age analysis

To study the age of genes belonging to different modules we dated the genes by their first appearance in the phylogeny. This consisted of retrieving the age of the oldest node of their Gene tree in Ensembl release 63 (Hubbard et al. 2009). Genes' age was described with one of the following categories: Fungi/Metazoa, Bilateria, Coelomata, Chordata, Eutelostomi, Clupeocephala, and *Danio rerio*. To fit the chi-square test requirements (more than 5 elements in a group) we merged the genes into five age categories: Fungi/Metazoa, Bilateria, Coelomata + Chordata, Eutelostomi, Clupeocephala + *Danio rerio*. Next, for every module we calculated the age distribution of its genes. We performed chi-square goodness of fit test to compare the observed and expected distributions of age classes in the modules. The expected distribution was estimated by classifying all zebrafish genes into one of the five age categories. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

## **Zebrafish-mouse orthologous genes**

Homology information of zebrafish and mouse genes was retrieved from Ensembl release 63 (Hubbard et al. 2009), using BioMart (Smedley et al. 2009). A total of 17 482 pairs of zebrafish-mouse orthologous genes had expression information in the zebrafish microarray data (14 293 zebrafish genes and 11 322 mouse genes). Among them there were 6441 one-to-one orthologous pairs, 5048 one-to-many orthologous pairs, and 2993 many-to-many orthologous pairs. 2901 zebrafish genes showed no orthology relationship with mouse genome. From further analysis we excluded 99 “apparent-one-to-one” gene pairs. For every module we calculated the number of genes that were in one-to-one, one-to-many, many-to-many and no orthology relation to mouse genes. Next, we performed chi-square goodness of fit test to compare the observed and expected distributions of orthology classes in the modules. The expected distribution was estimated by classifying all zebrafish genes into one of the four orthology categories. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

## **Gene expression conservation**

To study expression conservation between zebrafish genes assigned to the modules and their mouse one-to-one orthologs, we used gene expression data for 2883 orthologous gene pairs (the limiting factor being the mapping to both mouse microarrays). For genes that were mapped by several probe sets we averaged their signal across the probe sets for both species. In order to compare gene expression between two species, we first calculated the mean expression for zebrafish genes present in the modules and their one-to-one mouse orthologs. Due to the incompatibility of two mouse microarray data used it was difficult to provide a meaningful comparison of expression for the two species. To calculate the correlation between expression profiles between zebrafish and mouse we reduced their expression profiles to six metastages: zygote, cleavage, blastula, neurula, organogenesis, and post-embryonic stage (see Bastian et al. 2008 for detailed definition of metastage). For every module and every metastage we calculated the mean expression level for zebrafish genes and their mouse one-to-one orthologs, and next we calculated the Pearson correlation coefficient between them.

## Highly conserved non-coding elements

Location data for highly conserved non-coding elements (HCNE) between zebrafish and mouse (70% of identity) was retrieved from Ancora (Engström et al. 2008) ([http://ancora.genereg.net/downloads/danRer7/vs\\_mous](http://ancora.genereg.net/downloads/danRer7/vs_mous)). The file *HCNE\_danRer7\_mm9\_70pc\_50col.bed.gz* was downloaded and used in the analysis. For each of the 14 293 Ensembl genes considered in our analysis, we calculated the number of HCNE in regions of 500 base pairs upstream from the transcription start site. Next, for every module we performed a hypergeometric test to assess if they were significantly enriched in genes with HCNE. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level. In additional analyses, we calculated the number of HCNE in regions of 200 and 1000 base pairs upstream from the transcription start site, as well as in introns. Also, we repeated the analysis with HCNEs of 90% identity (see Supplementary Materials).

## Transposon-free regions

Location data for transposon-free regions (TFRs) in zebrafish was retrieved from Simons et al. (2007) (<http://www.biomedcentral.com/content/supplementary/1471-2164-8-470-S1.txt>). First, each TFR was associated with Ensembl ID of its closest transcript from genome assembly Zv6. Then for each Ensembl transcript ID we retrieved an Ensembl gene ID from genome assembly Zv9 (Ensembl release 63; Hubbard et al. 2009). For every module we performed a hypergeometric test to assess if they were significantly enriched in genes with TFRs in their proximity. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

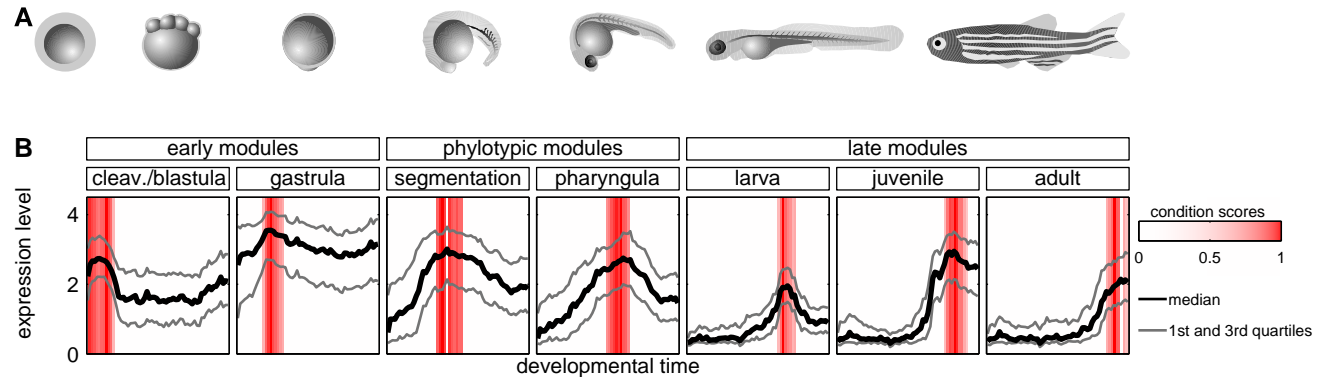
## Transcription factors

The set of transcription factors was defined based on GO category annotation: GO: 0006355, regulation of transcription, DNA-dependent. Among 14 293 Ensembl genes, 957 were annotated as transcription factors. For every module we performed a hypergeometric test to assess if they were significantly enriched in TFs. Next, we performed a hypergeometric test to assess if the TFs present in the modules were enriched in HCNEs and TFRs. To correct for multiple testing we applied the Bonferroni correction. We used 0.01 as a significance level.

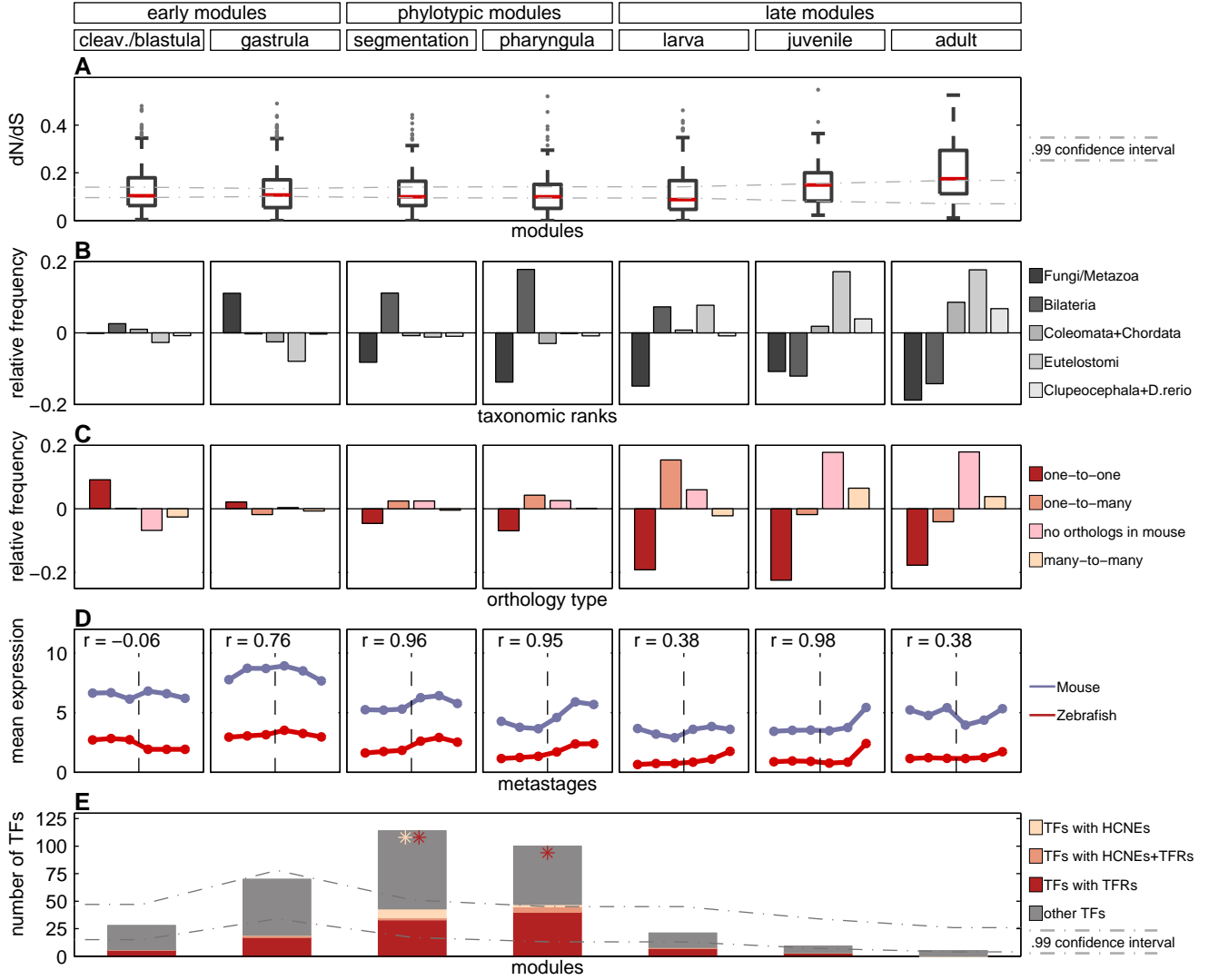
## Acknowledgments

We thank Tim Hohm, Anna Kostikova, Zoltan Kutalik, Eyal Privman, and Pavan Ramdya for useful comments on the manuscript. We thank Julien Roux and all members of MRR and SB labs for helpful discussion. We acknowledge the funding from Etat de Vaud, and Swiss National Science Foundation (ProDoc grant 1206624/1). SB was supported by the Swiss National Science Foundation (grant 31003A 130691/1) and the Swiss Institute of Bioinformatics. MRR was supported by the Swiss National Science Foundation (grant 31003A 133011/1) and the Swiss Institute of Bioinformatics.

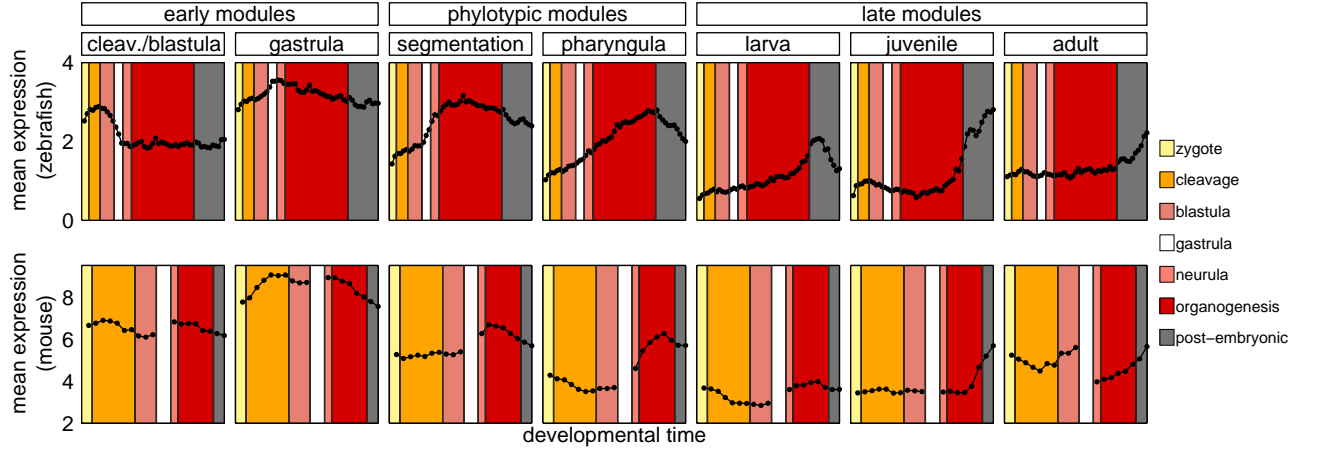
## Figures



**Figure 2. Modules of genes with time-specific expression during zebrafish development.** A) Zebrafish ontogeny (drawings of the embryos are based upon sketches and photographs from Kimmel et al. (1995)). B) Median, 25th and 75th percentiles of expression value of genes in modules. Red bars denote the condition scores assigned to developmental points by the ISA.



**Figure 3. Measures of developmental constraints for various gene properties.** A) Box and Whisker plot showing non-synonymous to synonymous substitution ratios ( $d_N/d_S$ ) for genes in the modules. The dash-dotted lines denote confidence interval for the median. B) Observed minus expected age distribution of genes in modules. C) Observed minus expected distribution of orthology type (between zebrafish and mouse) for genes in modules. D) Mean expression level of zebrafish genes in modules, and their one-to-one orthologs in mouse in six developmental metastages. The transition between the two mouse data sets is denoted with the vertical dashed line. The Pearson's correlation coefficients for zebrafish and mouse expression profiles are reported for every module. E) The number of transcription factors (TFs) in modules (whole bar) and their enrichment in highly conserved non-coding elements (HCNEs) and transposon-free regions (TFRs). The stars denote significant enrichment ( $p < 0.01$ ) of TFs in HCNEs (yellow) and in TFRs (red). The dash-dotted lines denote confidence interval for the expected number of TFs in modules.



**Figure 4. Developmental metastages.** Mean expression level of zebrafish genes in modules, and of their one-to-one orthologs in mouse. The same colors denote corresponding developmental metastages in zebrafish and mouse.

## References

- Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SGP, Lim AYM, Hajan HS, Collas P et al. 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res* **21**: 1328–38.
- Alexa A, Rahnenfuhrer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.
- Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, Robinson-Rechavi M. 2008. Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In *Data Integration in the Life Sciences*, editors A Bairoch, S Cohen-Boulakia, C Froidevaux, volume 5109 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 124–131.
- Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol* **24**: 26–53.
- Bergmann S, Ihmels J, Barkai N. 2003. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**: 031902.
- Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3**: e245.
- Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**: 25–36.
- Comte A, Roux J, Robinson-Rechavi M. 2010. Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evolution & development* **12**: 144–156.
- Demuth JP, Hahn MW. 2009. The life and death of gene families. *Bioessays* **31**: 29–39.
- Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**: 815–8.
- Duboule D. 1994. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl* : 135–42.
- Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–10.



- Elinson R. 1987. Change in Developmental Patterns: Embryos of Amphibians with Large Eggs. In *Development as an Evolutionary Process*, editor RE Raff RA, New York: Alan R. Liss., pp. 1–21.
- Engström PG, Fredman D, Lenhard B. 2008. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol* **9**: R34.
- Hartley RS, Rempel RE, Maller JL. 1996. In vivo regulation of the early embryonic cell cycle in *Xenopus*. *Dev Biol* **173**: 408–19.
- Hazkani-Covo E, Wool D, Graur D. 2005. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer’s third law. *J Exp Zool B Mol Dev Evol* **304**: 150–8.
- Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P et al. 2009. Ensembl 2009. *Nucleic Acids Res* **37**: D690–7.
- Ihmels J, Bergmann S, Barkai N. 2004. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**: 1993–2003.
- Irie N, Kuratani S. 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun* **2**: 248.
- Irie N, Sehara-Fujisawa A. 2007. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol* **5**: 1.
- Irimia M, Tena JJ, Alexis M, Fernandez-Miñan A, Maeso I, Bogdanovic O, de la Calle-Mustienes E, Roy SW, Gomez-Skarmeta JL, Fraser HB. 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res* .
- Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* **345**: 119–126.
- Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin EV. 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol* **21**: 2058–70.
- Kalinka A, Tomancak P. 2012. The evolution of early animal embryos: conservation or divergence? *Trends in Ecology & Evolution* **27**: 385–393.

- Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**: 811–4.
- Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn* **203**: 253–310.
- King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–16.
- Krumlauf R. 1994. Hox genes in vertebrate development. *Cell* **78**: 191–201.
- McDonald JH. 2009. *Handbook of Biological Statistics (2nd ed.)*. Sparky House Publishing, Baltimore, Maryland.
- Nei M. 2007. The new mutation theory of phenotypic evolution. *Proc Natl Acad Sci U S A* **104**: 12235–42.
- Ohno S et al. 1970. *Evolution by gene duplication*. London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.
- Poe S, Wake MH. 2004. Quantitative tests of general models for the evolution of development. *Am Nat* **164**: 415–22.
- Preuss TM, Cáceres M, Oldham MC, Geschwind DH. 2004. Human brain evolution: insights from microarrays. *Nat Rev Genet* **5**: 850–60.
- Prud’homme B, Gompel N. 2010. Evolutionary biology: Genomic hourglass. *Nature* **468**: 768–9.
- Quint M, Drost HG, Gabel A, Ullrich KK, Bönn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* .
- Raff RA. 1996. *The shape of life: genes, development, and the evolution of animal form*. Chicago; London: University of Chicago Press.
- Roux J, Robinson-Rechavi M. 2008. Developmental constraints on vertebrate genome evolution. *PLoS Genet* **4**: e1000311.
- Sandelin A, Bailey P, Bruce S, Engström PG, Klos JM, Wasserman WW, Ericson J, Lenhard B. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**: 99.

- Sander K. 1983. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In *Development and evolution*, editor WC Goodwin BC Holder N, Cambridge University Press, pp. 137–159.
- Seidel F. 1960. Körpergrundgestalt und Keimstruktur. Eine Erörterung über die Grundlagen der vergleichenden und experimentellen Embryologie und deren Gültigkeit bei phylogenetischen Überlegungen. *Zool Anz* **164**: 245–305.
- Simons C, Makunin IV, Pheasant M, Mattick JS. 2007. Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics* **8**: 470.
- Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. 2009. BioMart – biological queries made easy. *BMC Genomics* **10**: 22.
- Speed T. 2000. *Always log spot intensities and ratios*. URL <http://www.stat.berkeley.edu/users/terry/zarray/Html/>
- Stern DL. 2000. Evolutionary developmental biology and the problem of variation. *Evolution* **54**: 1079–91.
- Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol* **8**: R15.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**: 327–35.
- von Baer KE. 1828. *Ueber Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion*. Königsberg: Bornträger.
- Wang QT, Piotrowska K, Ciemerych MA, Milenkovic L, Scott MP, Davis RW, Zernicka-Goetz M. 2004. A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev Cell* **6**: 133–44.
- Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS Biol* **4**: e52.
- Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**: e7.

- Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**: 206–16.
- Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* **8**: 15–24.
- Yarden A, Geiger B. 1996. Zebrafish cyclin E regulation during early embryogenesis. *Dev Dyn* **206**: 1–11.
- Zhang J. 2003. Evolution by gene duplication - an update. *Trends Ecol Evol* **18**: 292–298.

# Supplementary Materials: The hourglass and the early conservation models — co-existing evolutionary patterns in vertebrate development

Barbara Piasecka<sup>1,2,4</sup>, Paweł Lichocki<sup>3</sup>, Sven Bergmann<sup>2,4,#</sup>, Marc Robinson-Rechavi<sup>1,4,#,\*</sup>

**1** Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

**2** Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland

**3** Laboratory of Intelligent Systems, EPFL, Lausanne, Switzerland

**4** Swiss Institute of Bioinformatics, Lausanne, Switzerland

# These authors contributed equally to this work.

\* E-mail: marc.robinson-rechavi@unil.ch

## Re-analysis of previous studies

### Domazet-Lošo and Tautz (2010)

In a recent paper, Domazet-Lošo and Tautz [1] suggested that *‘the phylotypic stage does express the oldest transcriptome set and that younger sets are expressed during early and late development’*. To study the relationship between gene expression, ontogeny and phylogeny, the authors proposed a measure called the “transcriptome age index” (TAI). In the main text (Box 1) we show that the transcriptome age measured with TAI [1] differs strongly if the log10-transformation of the data is applied. Here, we first discuss the advantages of log-transformation, and next we show that also applying several other measures and transformations of the data never reproduces the results reported in [1]. On the contrary, we find always that the age of the transcriptome decreases during development.

The microarray signal intensity values that were used in [1] display a log-normal distribution and span from 1 to  $10^5$  (figure S1). If one uses non-transformed data to calculate TAI, then the five orders of magnitude of difference between expressions of highly and lowly expressed genes translates into five orders of magnitude of difference of the weights of the phylogenetic ranks. In practice, this means that highly expressed genes are given a very high importance, whereas lowly expressed genes are given almost none (figure S2). It is disputable whether this is a correct interpretation of the biological reality, because even lowly expressed genes (which are a large majority) do play a role in the development and are shaped by evolutionary forces. Thus, one should not neglect them, if one wishes to interpret the TAI profile in the context of evolutionary constraints or evolutionary adaptation on the whole transcriptome, as in [1]. It can also be legitimate to study only a subset of genes, but then this should be done explicitly, and the properties of this subset should be well defined. In order to take into account all genes having a function during development, the data must be transformed, so that the weights of the phylogenetic ranks span a more comparable range. Of note, X-fold difference in signal intensity does not necessarily imply X-fold difference in RNA concentration [2].

Moreover, non-log-transformed data are very sensitive to outliers. We identified the probe A\_15\_P161596 as an outlier (figure S3A) which strongly distorts the TAI profile reported in [1]. If this single outlier is removed, a TAI peak during gastrulation – which in [1] was given an evolutionary interpretation and linked to the action of the group of genes that emerged in Metazoa – disappears and leaves the gastrulation trend less marked (figure S3B). In contrast, the presence of the outlier has little, if any, influence on the TAI profile calculated on log-transformed data (figure S3C), showing how the log-transformation leads to a more robust analysis.

Also, in [1], the authors used all 16 188 probes to calculate TAI. Since some of them map to the same gene, this results in signal multiplication for some phylostrata. To overcome this problem, we calculated TAI on data with averaged signal from probes mapped to the same gene. This changes the TAI pattern observed by the authors [1]: the oldest transcriptome now seem to be expressed in mid-larval stage, instead of the phylotypic stage (figure S4A). In contrast, the TAI profile calculated on log-transformed

data is more robust, as the pattern remains unchanged and does not depend on mapping to probes or genes (figure S4B).

Another approach to reduce the effect of highly expressed genes is to treat all expressed genes as equally important, i.e., recode as present-absent. This recovers the same pattern as log-transformation (figure S5). Of note, this approach was suggested in [1], without discussion of the results.

Finally, we searched for alternative measures of the evolutionary age of the transcriptome over ontogeny. We computed: (i) the difference in median expression profile of old genes vs. young genes (figure S6A) (similar to [3]); and (ii) the mean age of expressed genes (figure S6B). Both measures recover the decreasing trend over ontogeny. Moreover, measure (i) confirms that the male transcriptome is younger than the female one, consistent with the known fast evolution of male-specific genes [4], whereas the original analysis [1] indicated the opposite - younger female transcriptome.

Overall, it seems that the transcriptomic hourglass pattern reported previously [1] is not robust to different methods of analysis.

## Irie and Kuratani (2011)

Another analysis suggested that expression diverges less between vertebrate species in the phylotypic stage [5]. The authors calculated Spearman correlations between expression profiles of genes of four species: mouse, zebrafish, chicken and frog. They calculated these correlations for all possible pairs of stages, because it was not obvious how to map developmental stages between species. The correlations between expression profiles of genes were reported to be strongest on average at mid-development, supporting the hourglass model.

Here, we reproduced these results for three species: mouse, zebrafish and chicken. We did not re-analyze the frog data, because the expression was measured for tetraploid *Xenopus laevis*, whereas genome annotation available in Ensembl comes from diploid *Xenopus tropicalis*.

We first divided the development of three species into three general stages: early, middle and late (figure S7). The middle stage contained the time points from the phylotypic stage. The early stage contained the time points preceding the phylotypic stage. And, the late stage contained the time points following the phylotypic stage. We excluded from the analysis the first time point of mouse and zebrafish development, as they had no corresponding time point in the chicken development.

We verified if the middle stage displayed a higher expression similarity than the early stage. To this aim, for each pair of species, we compared the Spearman correlation values between all time points from the early stages of the two species with the Spearman correlation values between all time points from the middle stages of the two species (field A vs. field B on figure S7). We detected a statistically significant difference only for mouse and chicken (Mann-Whitney U test,  $p = 0.018$ ). However, because the time points from mid and late mouse stages displayed high correlation with almost any chicken time point, we performed a randomization test to confirm the significance of our observation. We permuted the order of chicken time points and compared again the correlation values between early and middle stages. Notably, among the 100 randomizations as many as 43 comparisons had P-value lower than the previously observed  $p = 0.018$ . Overall, the pattern of presumably conserved gene expression in middle development, reported in [5], was not significant for any pair of species.

## Artificial expression profiles for the ISA

We initialized the ISA with seven artificial expression profiles corresponding to consecutive developmental stages. Our main goal was to compare genes expressed in early, mid and late development. The early genes are known to divide into maternal genes (pre-MBT) and zygotic genes (post-MBT) [6]. Consequently, we originally envisioned four artificial expression profiles: pre-MBT, post-MBT, middle and late. During the ISA run, these profiles resulted in four modules containing genes with expression limited to

cleavage/blastula, gastrula, segmentation and juvenile stages, respectively. To cover the entire development we added three other artificial profiles corresponding to the missing stages (pharyngula, larva and adult) and we run ISA again. The seven profiles used to run the ISA are shown on the figure S8.

## Sequence conservation between mouse and human

In the main text, we investigated the conservation level of sequences of protein-coding genes in fishes. Here, we repeated this analysis by projecting the genes expressed in the seven modules to mouse-human orthologs. The orthology relationships, and the  $d_N$  and  $d_S$  values were obtained from Ensembl version 63 [7]. We retrieved 6039 zebrafish genes with one-to-one orthologs in mouse and human (the estimated divergence time is 61.5 MYA between the two mammalian species and 416 MYA with *Danio rerio* [8]) and the pairwise  $d_N/d_S$  between mouse and human genes using Biomart [9]. Other settings and the statistical analysis were the same as in the main text (see Methods). We found a good agreement between results reported in the main text and for mouse-human orthologs (compare figure 3A from the main text with figure S9).

## Highly conserved non-coding elements

We tested the sensitivity of the observed enrichment of HCNEs for genes expressed in mid-development, reported in the main text. To this aim, for each of the 14 293 Ensembl genes considered in our analysis, we calculated the number of HCNEs (70% identity) in regions of 200, and 1000 base pairs upstream from the transcription start site (TSS), as well as in the intronic regions. Also, we repeated the analysis looking for HCNEs in regions of 500 bp upstream from the transcription start site (as in the main text), but for HCNEs of 90% identity. To this aim we downloaded and used the file *HCNE\_danRer7\_mm9\_90pc\_50col.bed.gz*. Other settings and the statistical analysis were the same as in the main text (see Methods). The results of all four additional analyses are in a good agreement with the results reported in the main text (table S1).

**Table S1. P-values from HCNE enrichment analyses.**

	200bp	<b>500bp</b>	1000bp	intron	500bp(90%)
segmentation module	4.0e-3	<b>8.0e-6</b>	2.2e-7	2.5e-5	7.9e-1
pharyngula module	1.4e-3	<b>1.1e-4</b>	2.3e-7	2.0e-4	6.5e-4

The column in bold corresponds to the case reported in the main text.

## Microsynteny conservation

We checked for modules' enrichment in genes belonging to conserved ancestral microsyntenic pairs (CAMPs) [10]. From the list of 260 zebrafish CAMPs (Irimia, private communication) we selected 75 gene pairs involved in developmental regulation, i.e., "bystander gene + trans-dev gene". Both, bystander and trans-dev genes were reported to have conserved introns sequences. Thus, the trans-dev genes could potentially overlap with genes for which we detected enrichment in HCNEs in introns, as well as in the regions 1000 bp upstream from the TSS (CAMPs were shown to have very short intergenic regions, in some cases < 1kb). We crossed the list of trans-dev genes with the list of genes from each module. We performed hypergeometric test to assess if the overlap between genes was significant. To correct

for multiple testing we applied the Bonferroni correction. The number of CAMP-trans-dev genes in the seven modules were the following: 1, *n.s.*; 6, *n.s.*; 7,  $p = 0.018$ ; 4, *n.s.*; 2, *n.s.*; 1, *n.s.*; 0, *n.s.* The overrepresentation of trans-dev genes in the segmentation module stays in agreement with enrichment in HCNE detected in introns and in regions 1000 bp upstream from the TSS for genes belonging to this module. We also checked for enrichment in the remaining 185 CAMPs. Although they were reported to often be co-expressed, we did not find any such pair in our modules.

## References

1. Domazet-Lošo T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468: 815-8.
2. Kahn K (2008) Tutorial: Introduction to DNA Microarrays. [http://www.chem.ucsb.edu/~kalju/chem162/public/genechip\\_intro.html](http://www.chem.ucsb.edu/~kalju/chem162/public/genechip_intro.html).
3. Roux J, Robinson-Rechavi M (2008) Developmental constraints on vertebrate genome evolution. *PLoS Genet* 4: e1000311.
4. Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics* 8: 689-698.
5. Irie N, Kuratani S (2011) Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun* 2: 248.
6. Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, et al. (2011) Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Res* 21: 1328-38.
7. Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. *Nucleic Acids Res* 37: D690-7.
8. Benton MJ, Donoghue PCJ (2007) Paleontological evidence to date the tree of life. *Mol Biol Evol* 24: 26-53.
9. Smedley D, Haider S, Ballester B, Holland R, London D, et al. (2009) BioMart – biological queries made easy. *BMC Genomics* 10: 22.
10. Irimia M, Tena JJ, Alexis M, Fernandez-Miñan A, Maeso I, et al. (2012) Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res* .
11. LeProust E (2008) Agilent's Microarray Platform: How High-Fidelity DNA Synthesis Maximizes the Dynamic Range of Gene Expression Measurements. Application Note - Agilent Technologies 5989-9159EN. [http://www.chem.agilent.com/en-US/Search/Library/\\_layouts/Agilent/PublicationSummary.aspx?whid=](http://www.chem.agilent.com/en-US/Search/Library/_layouts/Agilent/PublicationSummary.aspx?whid=)
12. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19: 327-35.
13. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A* 106: 7273-80.



## Figures

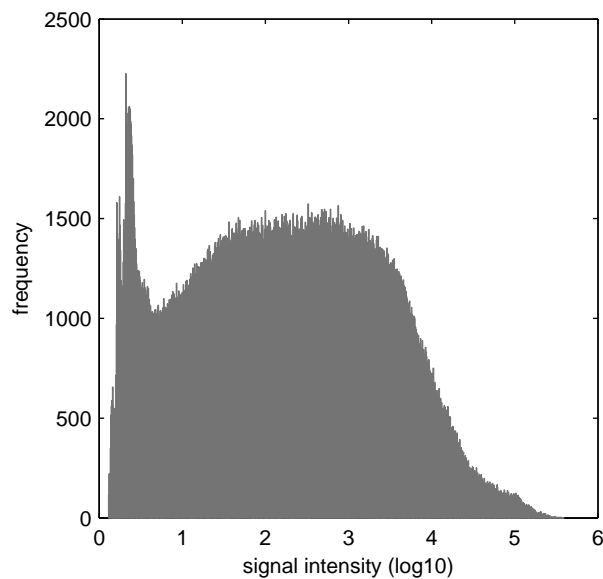


Figure S1. Total distribution of signal intensity from all 140 microarrays.

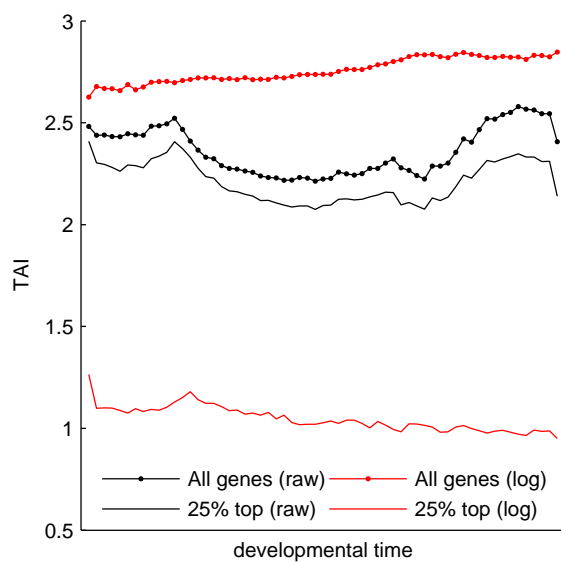
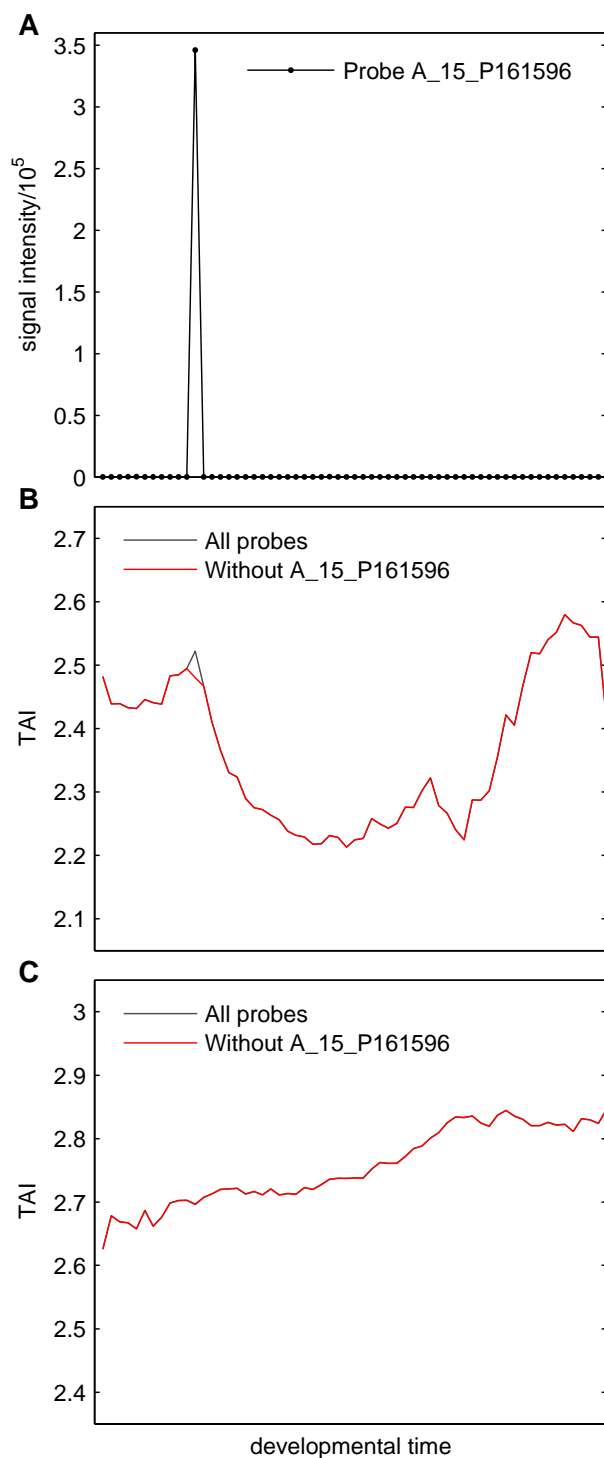
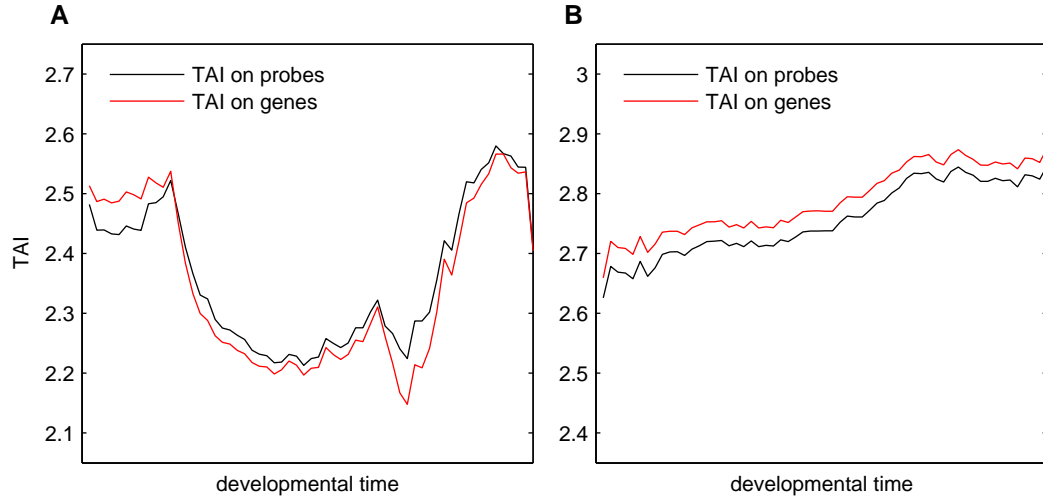


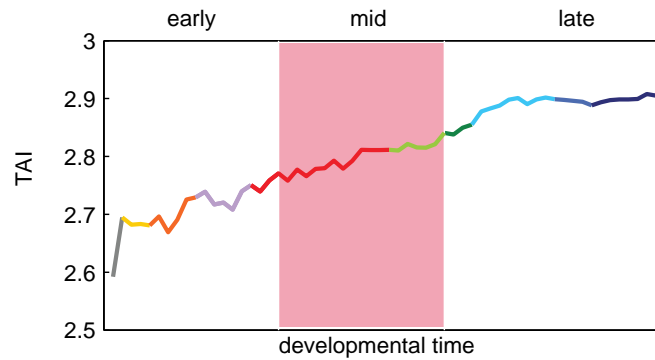
Figure S2. TAI hourglass pattern is driven by the subset of most expressed genes. TAI calculated using untransformed (black) and log10-transformed (red) gene expression intensities across zebrafish development. In both cases, TAI is calculated using the entire data sets (dotted line), or using the 25% highest partial concentrations<sup>1</sup> chosen separately for each stage (solid line).



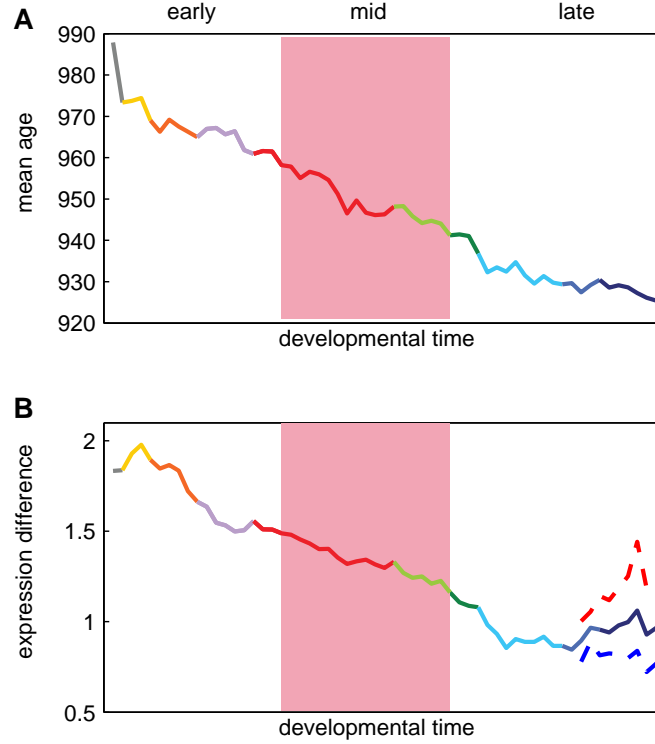
**Figure S3. Sensitivity to outliers.** (A) Raw expression signal of probe A\_15\_P161596 across zebrafish development. (B) TAI calculated on non-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey). (C) TAI calculated on log10-transformed data across zebrafish development without this probe (red) and the effect of this probe on TAI pattern (grey).



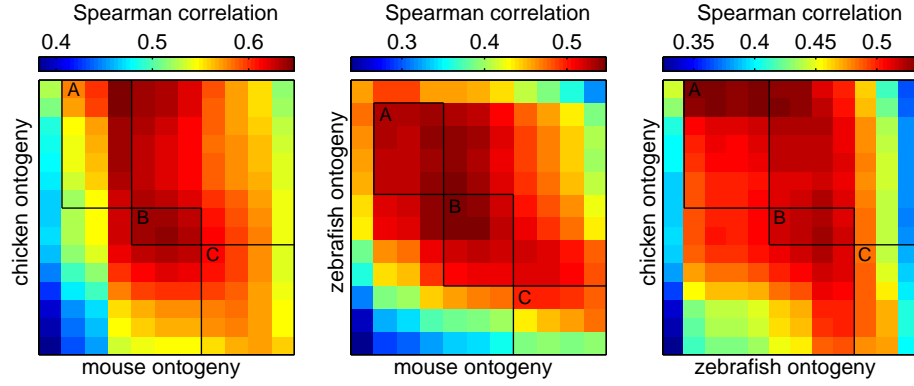
**Figure S4. TAI calculated using expression intensities of genes, instead of probes, across zebrafish development.** For each gene we averaged the signal intensity from all corresponding probes. After this process 16 188 probes' intensities values were reduced to 12 892 genes' intensities values, which were used to weight the phylogenetic ranks of genes (if two different phylostrata were assigned to the same gene, the older one was chosen). (A) non-transformed data was used. (B) log10-transformed data was used.



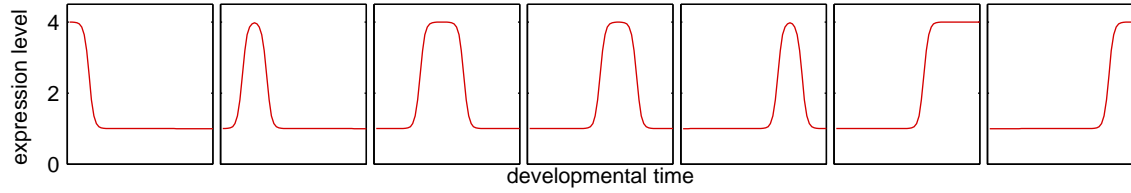
**Figure S5. TAI calculated using genes recoded as present-absent across zebrafish development.** At a given stage of development, if the log10-intensity value of a gene is above one [11], its expression is set to 1, otherwise it is set to 0. Other notations as in figure 1 (in main text).



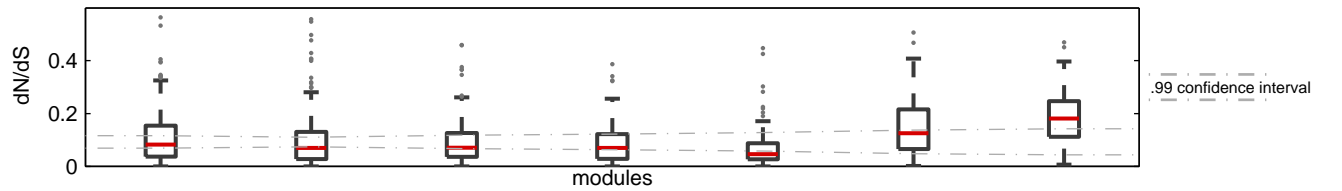
**Figure S6. Alternative measures of transcriptome age.** (A) Mean age of genes expressed across zebrafish development; age estimated with the TimeTree database ([www.timetree.org](http://www.timetree.org)). A gene is considered expressed at a given stage of development if its log10-intensity is above one [11]. (B) Difference between median expression profiles of old genes and young genes across zebrafish development. Here, the genes that have emerged before the evolution of Metazoa are considered old and the genes that have emerged since the ancestor of Euteleostomi are considered young. The difference between the two groups is always positive, reflecting that old genes tend to be more expressed than young genes [13]. The results are robust to the choice of cutoffs used to define old and young genes (data not shown). Red dashed line - female data, blue dashed line - male data. Other notations as in figure 1 (main text).



**Figure S7. Correlation between expression levels of genes across developmental time points of mouse, chicken and zebrafish.** Field A denotes the early stages, field B denotes the phylotypic stages, and field C denotes the late stages of development.



**Figure S8. Artificial expression profiles used to initialize the ISA:** pre-MBT, post-MBT, “middle”, pharyngula, larva, “late”, adult. These profiles resulted in modules containing genes expressed specifically in: cleavage/blastula, gastrula, segmentation, pharyngula, larva, juvenile, and adult, respectively.



**Figure S9.  $d_N/d_S$  ratio for human-mouse one-to-one orthologs.** The orthologs were obtained by projecting the genes expressed in the zebrafish modules to their one-to-one orthologs in mouse and human.

**Table S2.** The list of modules and their enriched GO categories (biological process).

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
<b>Module 1</b>						
GO:0006468	protein amino acid phosphorylation	446	29	13.7	1.00E-04	7.00E-04
GO:0090244	Wnt receptor signaling pathway involved in somitogenesis	2	2	0.06	9.40E-04	6.58E-03
GO:0006470	protein amino acid dephosphorylation	98	9	3.01	3.09E-03	2.16E-02
GO:0031290	retinal ganglion cell axon guidance	24	4	0.74	5.70E-03	3.99E-02
GO:0043149	stress fiber assembly	5	2	0.15	8.84E-03	6.19E-02
GO:0090090	negative regulation of canonical Wnt receptor signaling pathway	5	2	0.15	8.84E-03	6.19E-02
GO:0021915	neural tube development	31	4	0.95	1.43E-02	9.98E-02
GO:0042451	purine nucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
GO:0042455	ribonucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
GO:0046129	purine ribonucleoside biosynthetic process	8	2	0.25	2.33E-02	1.63E-01
<b>Module 2</b>						
GO:0016055	Wnt receptor signaling pathway	80	14	4.43	1.10E-04	7.70E-04
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	6	4	0.33	1.30E-04	9.10E-04
GO:0042664	negative regulation of endodermal cell fate specification	6	4	0.33	1.30E-04	9.10E-04
GO:0035468	positive regulation of signaling pathway	30	8	1.66	1.70E-04	1.19E-03
GO:0010159	specification of organ position	3	3	0.17	1.70E-04	1.19E-03
GO:0035050	embryonic heart tube development	70	21	3.88	2.00E-04	1.40E-03
GO:0001706	endoderm formation	18	9	1	2.80E-04	1.96E-03
GO:0060218	hemopoietic stem cell differentiation	7	4	0.39	2.90E-04	2.03E-03
GO:0007420	brain development	149	21	8.26	4.10E-04	2.87E-03
GO:0030903	notochord development	31	10	1.72	5.00E-04	3.50E-03
GO:0014028	notochord formation	4	3	0.22	6.50E-04	4.55E-03
GO:0001522	pseudouridine synthesis	9	4	0.5	9.40E-04	6.58E-03
GO:0045893	positive regulation of transcription, DNA-dependent	47	9	2.6	9.50E-04	6.65E-03
<b>Module 3</b>						
GO:0009952	anterior/posterior pattern formation	91	19	3.45	1.10E-04	7.70E-04
GO:0048741	skeletal muscle fiber development	14	5	0.53	1.10E-04	7.70E-04
GO:0030510	regulation of BMP signaling pathway	15	5	0.57	1.70E-04	1.19E-03
GO:0043049	otic placode formation	15	5	0.57	1.70E-04	1.19E-03
GO:0030901	midbrain development	15	5	0.57	1.70E-04	1.19E-03
GO:0021523	somatic motor neuron differentiation	4	3	0.15	2.10E-04	1.47E-03
GO:0042694	muscle cell fate specification	4	3	0.15	2.10E-04	1.47E-03
GO:0021508	floor plate formation	9	4	0.34	2.20E-04	1.54E-03
GO:0033334	fin morphogenesis	57	9	2.16	2.60E-04	1.82E-03
GO:0007156	homophilic cell adhesion	58	9	2.2	3.00E-04	2.10E-03
GO:0007517	muscle organ development	59	16	2.23	3.00E-04	2.10E-03
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	71	10	2.69	3.10E-04	2.17E-03
GO:0009888	tissue development	329	41	12.46	3.40E-04	2.38E-03
GO:0031016	pancreas development	39	7	1.48	5.70E-04	3.99E-03
GO:0030182	neuron differentiation	156	27	5.91	5.90E-04	4.13E-03
GO:0009953	dorsal/ventral pattern formation	65	9	2.46	7.10E-04	4.97E-03
GO:0007399	nervous system development	326	60	12.34	8.40E-04	5.88E-03

Continued on next page

Table S2 – continued from previous page

GO ID	Term	Annot.	Sign.	Expect.	elim p	bonf p
GO:0007223	Wnt receptor signaling pathway, calcium modulating pathway	21	5	0.8	9.30E-04	6.51E-03
GO:0021984	adenohypophysis development	9	5	0.34	9.70E-04	6.79E-03
GO:0001708	cell fate specification	36	9	1.36	1.06E-03	7.42E-03
<b>Module 4</b>						
GO:0030154	cell differentiation	422	39	13.18	1.20E-04	8.40E-04
GO:0030902	hindbrain development	57	11	1.78	2.30E-04	1.61E-03
GO:0050769	positive regulation of neurogenesis	12	4	0.37	3.80E-04	2.66E-03
GO:0048663	neuron fate commitment	13	4	0.41	5.30E-04	3.71E-03
GO:0048593	camera-type eye morphogenesis	47	9	1.47	8.60E-04	6.02E-03
GO:0030900	forebrain development	51	7	1.59	9.60E-04	6.72E-03
GO:0051091	positive regulation of transcription factor activity	7	3	0.22	9.60E-04	6.72E-03
GO:0030901	midbrain development	15	4	0.47	9.70E-04	6.79E-03
GO:0021915	neural tube development	31	5	0.97	2.50E-03	1.75E-02
GO:0002043	blood vessel endothelial cell proliferation involved in sprouting angiogenesis	3	2	0.09	2.86E-03	2.00E-02
<b>Module 5</b>						
GO:0007602	phototransduction	10	4	0.28	1.10E-04	7.70E-04
GO:0006813	potassium ion transport	79	9	2.18	3.10E-04	2.17E-03
GO:0018298	protein-chromophore linkage	13	4	0.36	3.40E-04	2.38E-03
GO:0007156	homophilic cell adhesion	58	7	1.6	1.03E-03	7.21E-03
GO:0006836	neurotransmitter transport	36	8	1	1.62E-03	1.13E-02
GO:0006814	sodium ion transport	52	6	1.44	2.96E-03	2.07E-02
GO:0007267	cell-cell signaling	41	8	1.13	3.14E-03	2.20E-02
GO:0007194	negative regulation of adenylate cyclase activity	5	2	0.14	7.21E-03	5.05E-02
GO:0007268	synaptic transmission	21	6	0.58	1.04E-02	7.25E-02
GO:0006208	pyrimidine base catabolic process	6	2	0.17	1.06E-02	7.43E-02
<b>Module 6</b>						
GO:0006805	xenobiotic metabolic process	3	2	0.05	9.20E-04	6.44E-03
GO:0006584	catecholamine metabolic process	6	2	0.11	4.44E-03	3.11E-02
GO:0019882	antigen processing and presentation	20	3	0.35	4.95E-03	3.47E-02
GO:0006022	aminoglycan metabolic process	22	3	0.39	6.52E-03	4.56E-02
GO:0046686	response to cadmium ion	8	2	0.14	8.10E-03	5.67E-02
GO:0009607	response to biotic stimulus	47	4	0.83	9.29E-03	6.50E-02
GO:0000272	polysaccharide catabolic process	9	2	0.16	1.03E-02	7.21E-02
GO:0006026	aminoglycan catabolic process	9	2	0.16	1.03E-02	7.21E-02
GO:0055114	oxidation reduction	409	14	7.23	1.35E-02	9.42E-02
GO:0006144	purine base metabolic process	11	2	0.19	1.54E-02	1.08E-01
<b>Module 7</b>						
GO:0043687	post-translational protein modification	748	16	8.26	7.70E-03	5.39E-02
GO:0050896	response to stimulus	622	16	6.87	9.40E-03	6.58E-02
GO:0051707	response to other organism	40	3	0.44	9.60E-03	6.72E-02
GO:0006950	response to stress	329	9	3.63	1.04E-02	7.28E-02
GO:0006508	proteolysis	391	10	4.32	1.10E-02	7.70E-02
GO:0051715	cytolysis of cells of another organism	1	1	0.01	1.10E-02	7.70E-02
GO:0044403	symbiosis, encompassing mutualism through parasitism	1	1	0.01	1.10E-02	7.70E-02

Continued on next page

**Table S2 – continued from previous page**

<b>GO ID</b>	<b>Term</b>	<b>Annot.</b>	<b>Sign.</b>	<b>Expect.</b>	<b>elim p</b>	<b>bonf p</b>
GO:0051801	cytolysis of cells in other organism involved in symbiotic interaction	1	1	0.01	1.10E-02	7.70E-02
GO:0031640	killing of cells of another organism	1	1	0.01	1.10E-02	7.70E-02
GO:0070193	synaptonemal complex organization	1	1	0.01	1.10E-02	7.70E-02

Annot. — total number of genes annotated with a given GO category; Sign. — number of (significant) genes in the module annotated with a given GO category; Expect. — expected number of genes in the module annotated with a given GO category; elim p — P-value from “elim” algorithm of topGO, bonf p — P-value after Bonferroni correction.